Prioritizing risk genes for neurodevelopmental disorders using pathway information

Author List:

Hoang T. Nguyen[1], Amanda Dobbyn[1,2], Alex Charney[1], Julien Bryois[3], Nathan G. Skene[4], Laura M. Huckins[1], Weiqing Wang[1], Douglas M Ruderfer[5], Xinyi Xu[6], Menachem Fromer[7], Shaun M Purcell[8], Matthijs Verhage[9], August B. Smit[10], Jens Hjerling-Leffler[4], Joseph Buxbaum[6], Dalila Pinto[6,11,12], Xin He[13], Patrick F Sullivan[14], Eli A. Stahl[1,15]

1. Supplementary Information

1.1 Analyzing of simulated data

1.1.1 Test on single gene sets

In general, when eGSs were used, gTADA prioritized more significant genes than extTADA for both true positive genes and all positive genes (Figure 1). The number of genes increased when larger gene sets were used. When small sample sizes were used, larger differences were observed for the PP threshold > 0.8. For random gene sets, the gene-count results were nearly equal between gTADA and extTADA (Figure S1).

To see the influence of gene-set sizes on PPs, the relationship between PPs and observed FDRs were also tested. For GSs less than 1000 genes, PPs > 0.8 were nearly equal to FDRs<0.1 (Figure S1). For large GSs (> 1000 genes), FDRs were slightly larger than 0.1 when PPs were > 0.8.

1.1.2 Type I error for calling enriched gene sets

We also calculated Type I error rates for calling a random GS as an enriched GS. We tested for two situations: p values < alpha thresholds and CIs > 0, and only p values < alpha thresholds. gTADA had a small inflated type I error at low α levels, but it correctly exhibited for medium and high α levels. When only p values were used to test GS enrichment, the errors were slightly higher than those of both p-value and CI information, but it was still well calibrated (Table S3). Only 1.6% of random GSs were called enriched GSs (Table S3).

1.1.3 Test for multiple gene sets

We also tested whether multiple GSs could increase the power of prioritizing genes for gTADA. We added GSs by using a forward-selection based strategy (Details in Methods). As expected, the number of prioritized genes increased when GS numbers increased (Figure S2); however, observed false discovery rates also increased.

1.2 Analyzing of real data.

1.2.1 Insight of the rare variant genetic architecture of EPI

We applied gTADA without gene sets to infer genetic parameters of EPI. The mean relative risks (MeanRRs) of de novo mutations were approximately 20. For case/control (CC) data, MeanRRs

of familial non-acquired focal epilepsy (familiar NAFE) and familial genetic generalized epilepsy, (familiar GGE) were nearly equal (5.1 and 5.2 respectively). Surprisingly, the mean RRs of the sporadic non-acquired focal epilepsy (NAFE) CC sample was 4.2 which was not much smaller than other two CC population samples. These CC results were much larger than the result of the Epi K. consortium and Epilepsy Phenome/Genome Project (2017). Therefore, we recalculated the CC ratios and the confident intervals of these ratios for the three population samples with different PP thresholds by using a bootstrapping approach. For all PP > 0.1, the CC ratios were larger than 4 (Table S14). These ratios were strongly significant when we compared with random GSs having the same size from the whole genes ($p < 9.9e-4$, Table S14). Therefore, all the genes with PP > 0.1 from gTADA without GS (2102 genes) could be considered as an enriched GS across three CC population samples. This supplied more information for EPI because significant differences between cases and controls were only reported for two familiar NAFE and GGE population samples in the study of Epi K. consortium and Epilepsy Phenome/Genome Project (2017).

To better understand the results of gene sets from gTADA, we also used the same method above to test for five gene sets (43 known EPI genes; FMRP, NMDAR, seizures and ion gene sets) used in the study of Epi K. consortium and Epilepsy Phenome/Genome Project (2017). Our results were similar to the Epi K. consortium and Epilepsy Phenome/Genome Project (2017) for the four gene sets. For 43 known EPI gene, highly significant results were observed ($p < 9.9-e4$) for familiar NAFE and familiar GGE samples, but the result for sporadic NAFE sample was not significant ($p \sim 0.18$). For four other gene sets, the most significant result was the ion gene set ($p < 6.9-e3$) in the familiar NAFE samples (Table S14).

2. Supplementary figures

Figure S1: The performance of gTADA in the prioritization of top genes for single gene sets (GSs). Left panel compares gene counts between extTADA and gTADA for different sample sizes. The left panel is for single gene sets in which random gene sets (rGSs) and enriched gene sets (eGSs) are presented side by side. These are gene counts with different posterior probabilities (PP) of 0.95 and 0.8. The right panel describes the correlation between PPs and observed false discovery rates (FDRs).

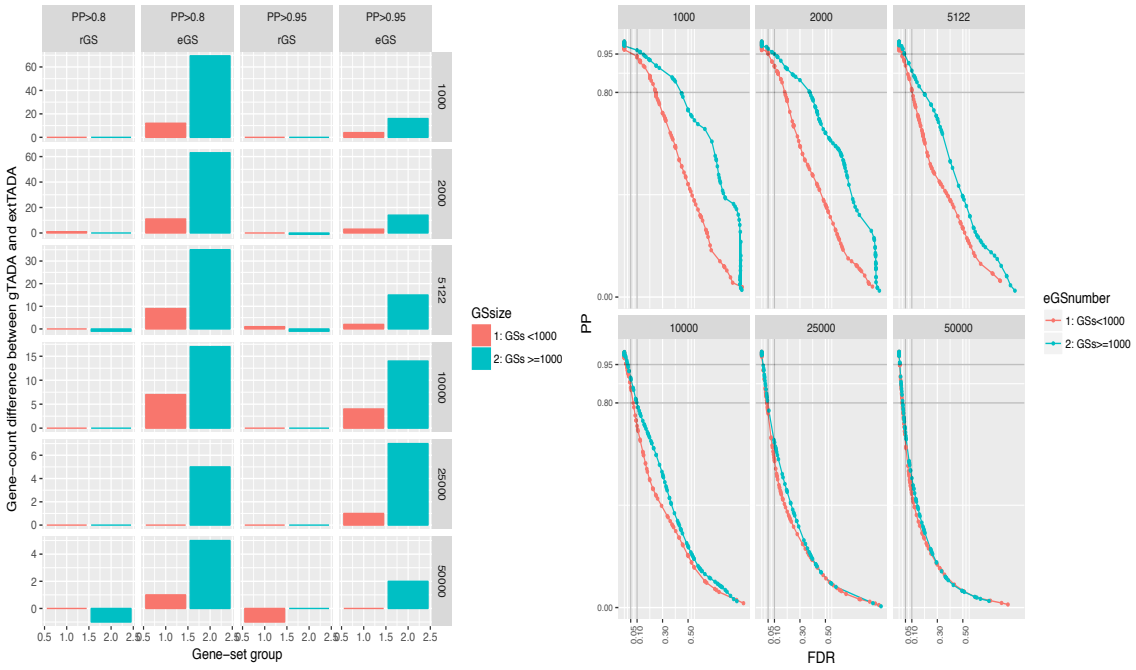


Figure S2: The performance of gTADA in the prioritization of top genes for multiple gene sets (mGSs) and for pooling the results of single GSs (pGSs). Left panel compares gene counts between extTADA and gTADA for different numbers of GSs: these are gene counts with different posterior probabilities (PP) of 0.95 and 0.8. The right panel describes the correlation between PPs and observed false discovery rates (FDRs) for mGSs.

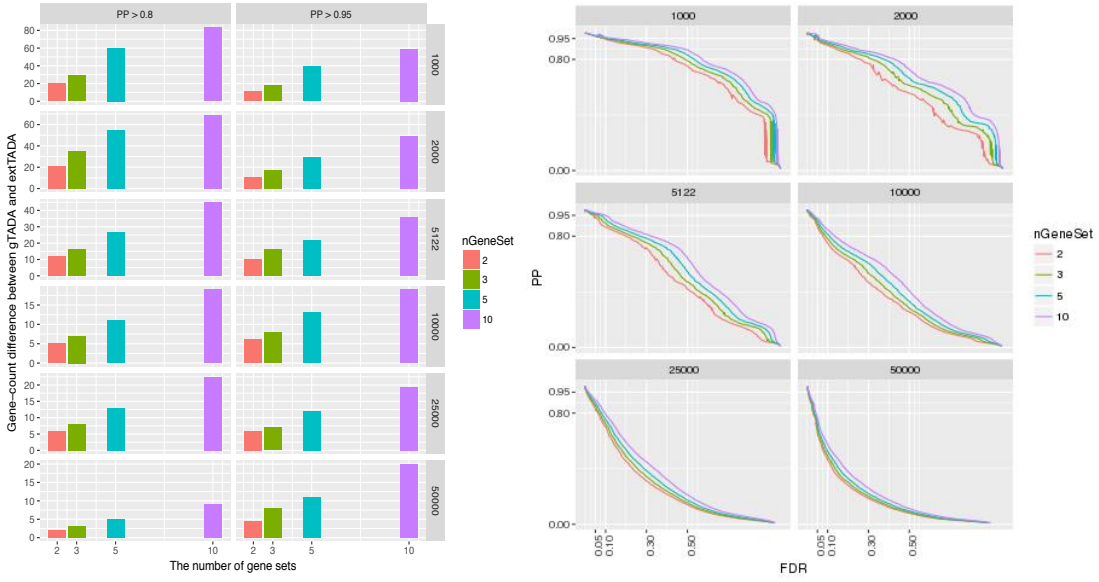


Figure S3: P-value correlation between gTADA and previous methods. These results are for 186 gene sets (GSs) analyzed in current study and in the previous study of our group. Left panels show correlations between gTADA and the two previous methods: permutation based method (PE) and posterior probability based method (PP). Right panels describe numbers of gene sets which are identified by three methods. PE used the top 500 genes with the smallest FDRs from extTADA to test the enrichment of the 186 GSs. PP calculated the sum of the posterior probabilities of a tested GS and compare the sum with those of random GS having the same size as the tested GS.

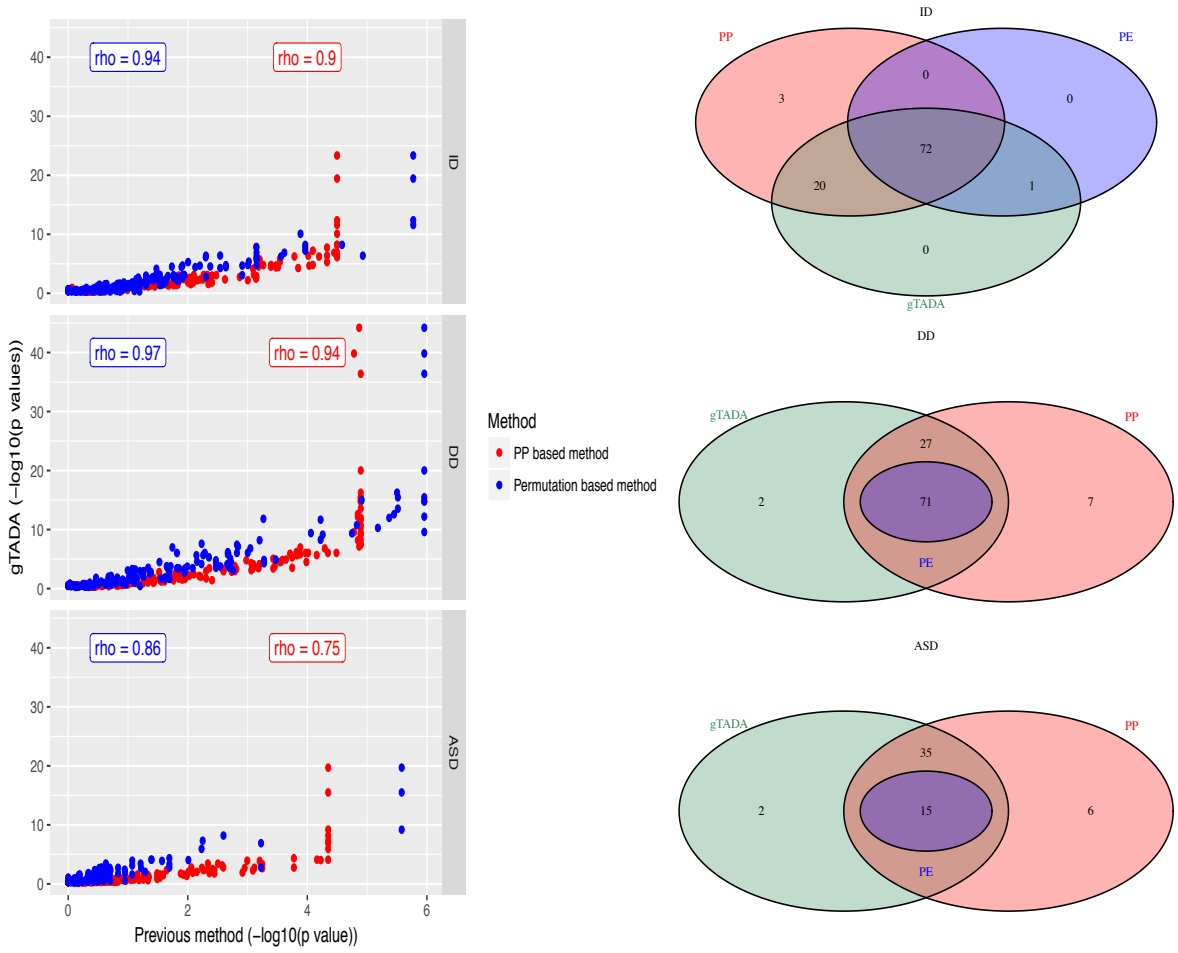


Figure S4: gTADA results for GTEx tissues. These are credible intervals (CIs) and modes estimated by gTADA for tissues. Red color intervals are for enriched tissues after adjusting for multiple tests.

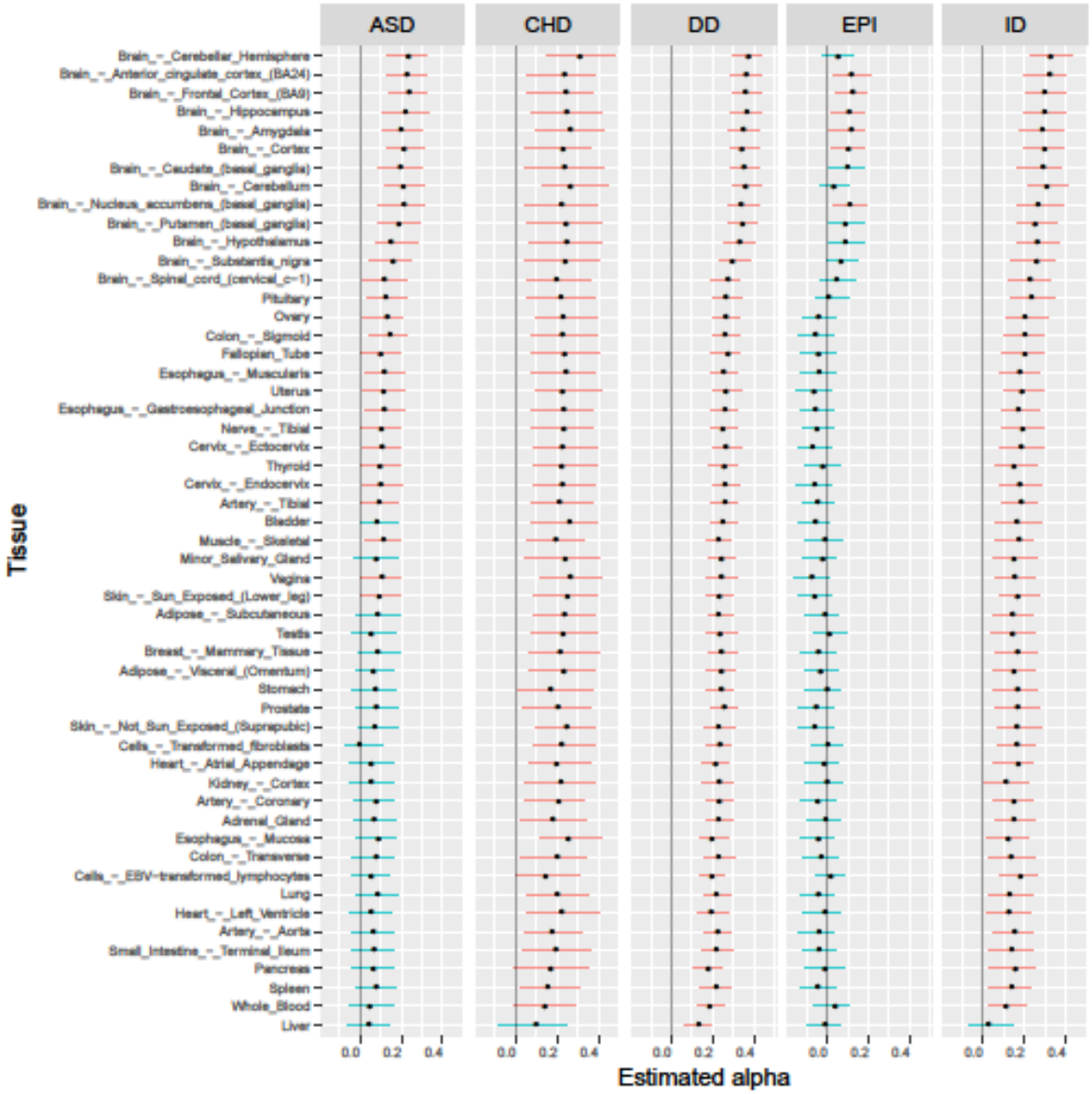


Figure S5: The number of overlapping genes between different gene sets and no GS (noGS) for EPI.

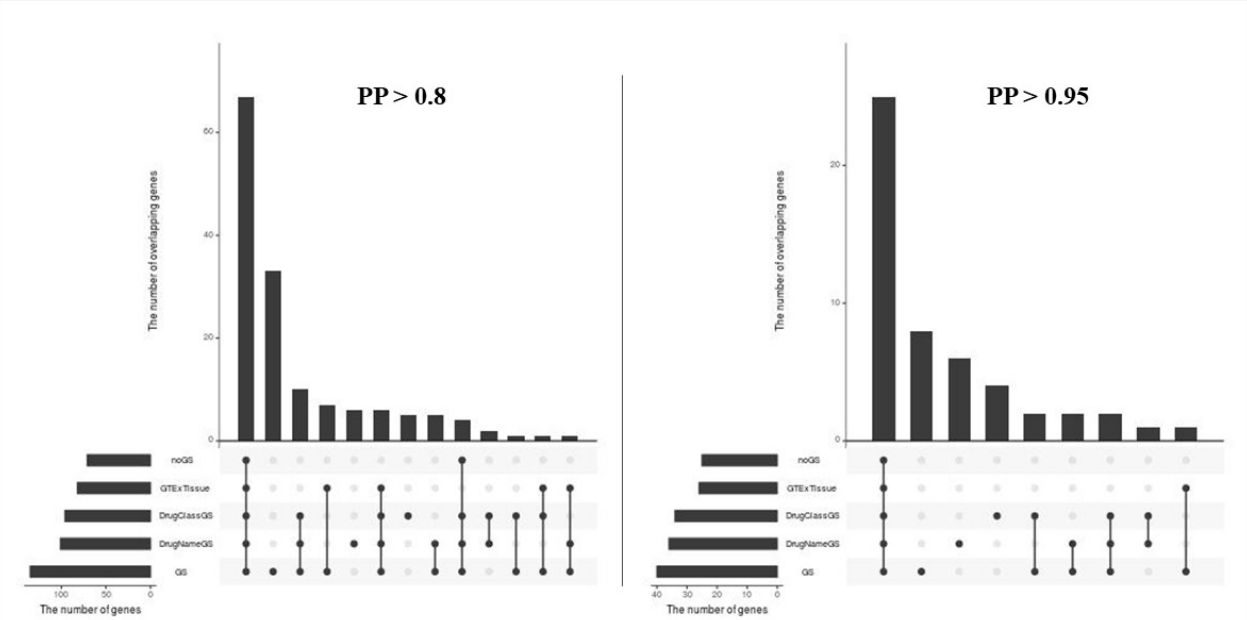


Figure S6: Correlation between the number of gene sets and observed false discovery rates (FDRs) by using different thresholds of posterior probabilities (PPs). These are simulation results for enriched gene sets of EPI. The genetic parameters of de novo mutations and rare case-control variants are from the analysis of 356 trios + 5,704 cases and controls.

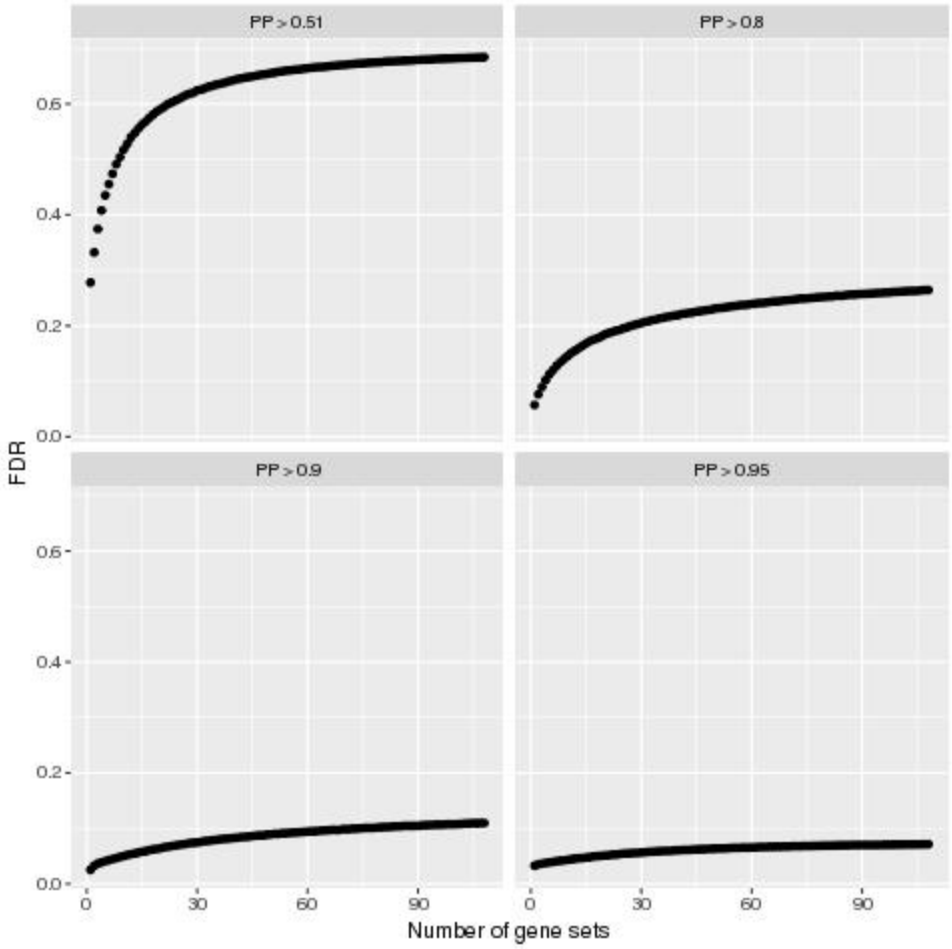


Figure S7: The enrichment results of single-cell RNA sequencing (scRNAseq) data in different communities. These results are for five communities generated by GeNets (Hu, et al., 2017). For each community, scRNAseq data were tested for genes from gTADA only and for all genes (gTADA genes and candidate genes) from GeNets. These two types of results are described as gTADA and PPI for each community.

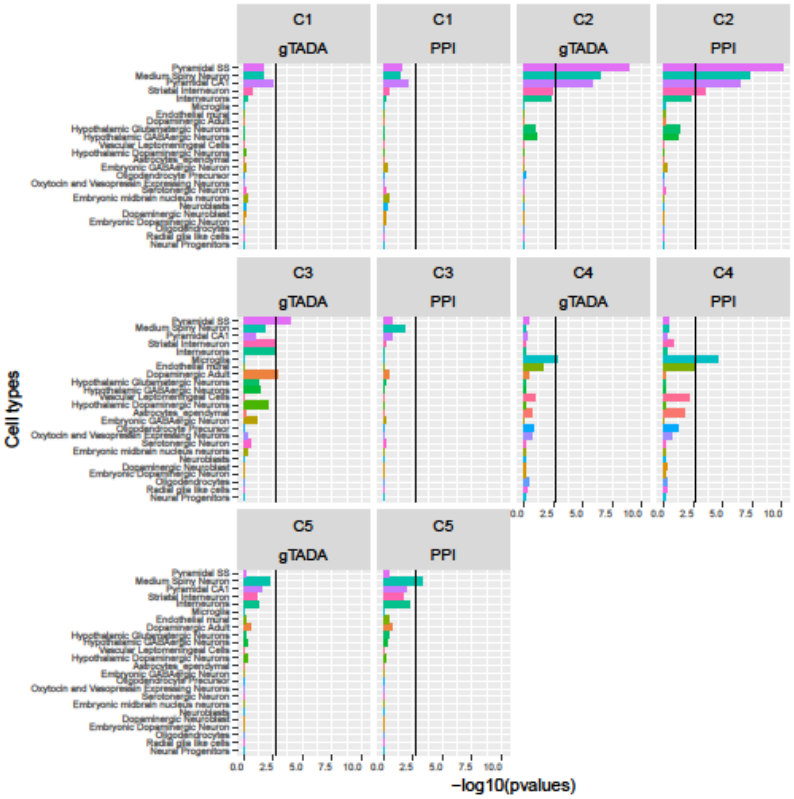


Figure S8: PPI-network analysis of the top prioritized genes (PP > 0.8) from gene sets and drug-target information.

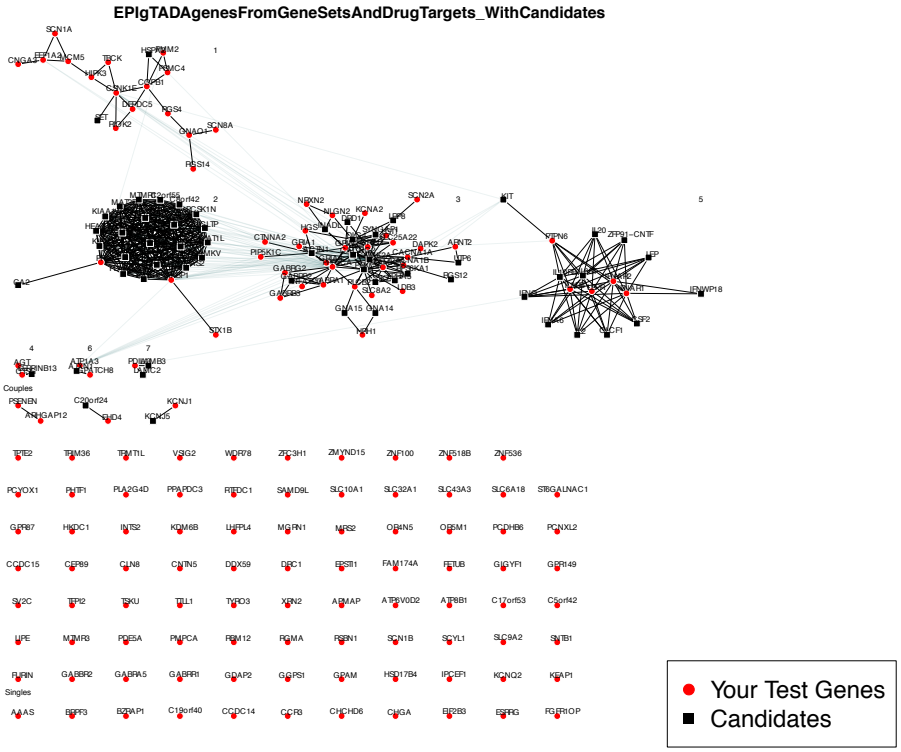


Figure S9: Spatiotemporal gene expression for prioritized genes from gTADA (PP > 0.8 from candidate gene sets).

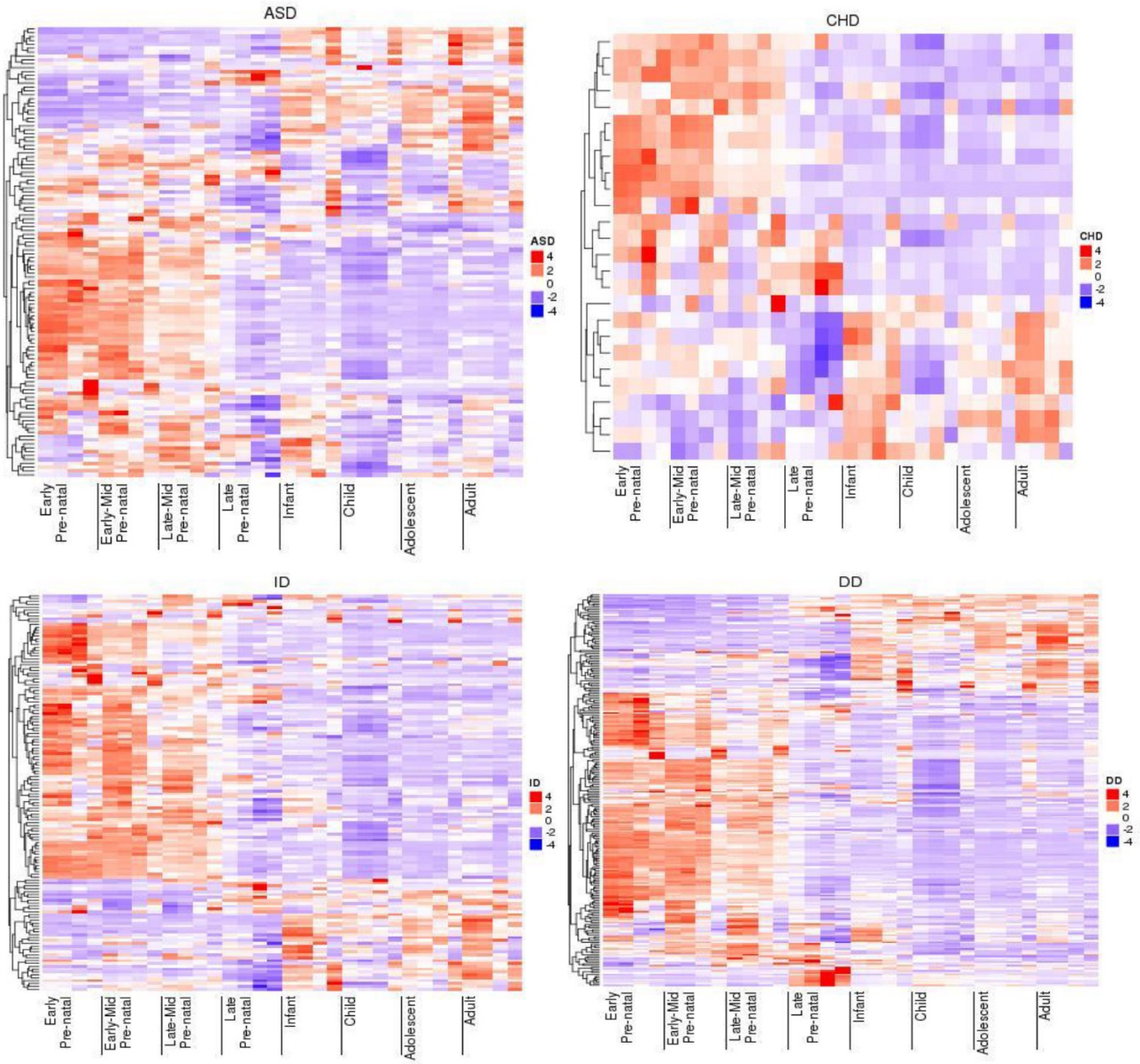
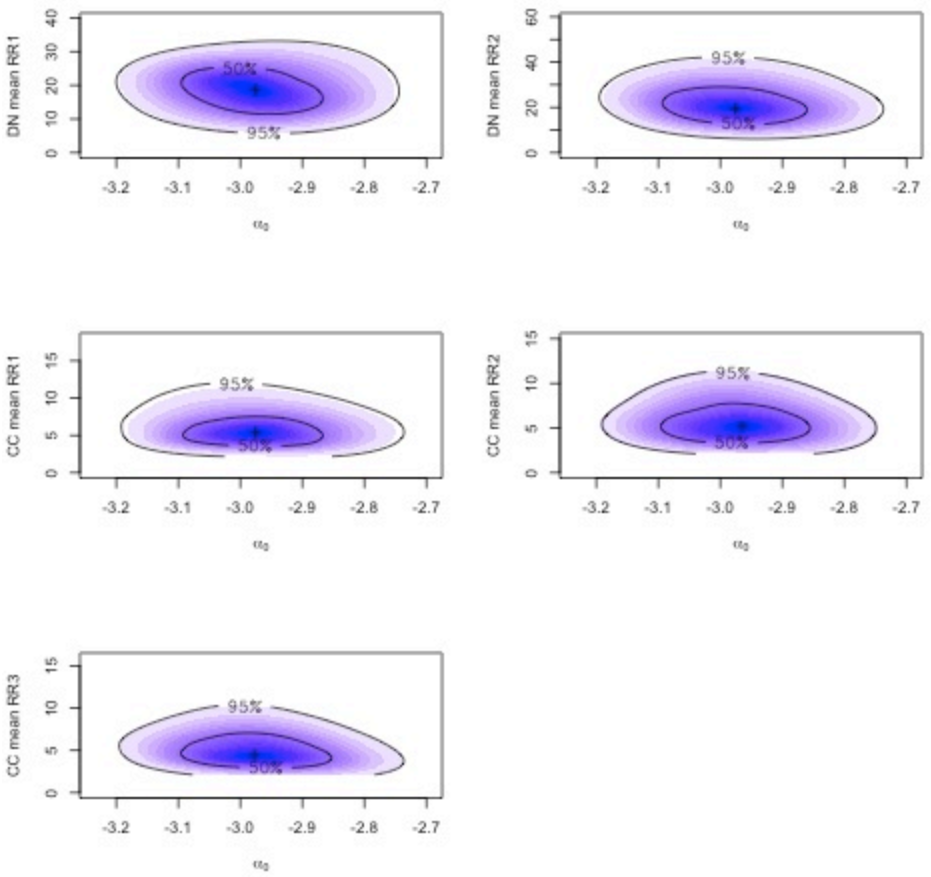


Figure S10: EPI genetic parameters from de novo (DN) and rare case-control (CC) data sets. Y axes are mean relative risks (mean RRs) for two DN classes, and three CC population samples. X axes are the intercept in the logistic regression: $\alpha_0 = ln\,(\frac{p_i}{1-p_i})$, p_i *is the probability of a gene being a risk gene.*



3. Supplementary Tables

Table S1: Parameters used in gTADA. Statistical models for de novo (dn) and case/control (cc) data are from Nguyen, et al. (2017). $N_{dn}, N_1 and\ N_0$ *are sample sizes for families, cases and controls respectively.* x_{dn}, x_1 *and* x_0 *are de novo, case and control counts in that order at a given* i^{th} *gene.*π_i *is the prior probability of being a risk gene for the* i^{th} *gene. K is the number of gene sets.* GS_{ij}*is the value of the* j^{th} *gene set at a given* i^{th} *gene.*

Data model/Equation	Parameter prior	Hyper prior
$x_{dn} \sim Poisson(2N_{dn}\mu\gamma_{dn})$	$\gamma_{dn} \sim Gamma(\bar{\gamma}_{dn} * \beta_{dn}, \beta_{dn})$ $\beta_{dn} = e^{a*\bar{\gamma}_{dn}^{b}+c}$	$\bar{\gamma}_{dn} \sim Gamma(\bar{\bar{\gamma}}_{dn}, \bar{\beta}\ )$
$x_{ca} \sim Poisson(2N_1 q\gamma_{cc})$	$\gamma_{cc} \sim Gamma(\bar{\gamma}_{cc} * \beta_{cc}, \beta_{cc})$ $\beta_{cc} = e^{a*\bar{\gamma}_{cc}^{b}+c}$ $q \sim Gamma(\rho, \nu)$	$\bar{\gamma}_{cc} \sim Gamma(\bar{\bar{\gamma}}_{cc}, \bar{\beta}_{cc})$ $\frac{\rho}{\nu} = mean(\sum(x_{cn} + x_{ca}))$ $\nu = 200$

$x_{cn} \sim Poisson(2N_0 q)$	$q \sim Gamma(\rho, \nu)$	$\frac{\rho}{\nu} = mean(\sum(x_{cn} + x_{ca}))$ $\nu = 200$
$\pi_i = \frac{e^{1+\sum_{j=1}^{K} \alpha_j GS_{ij}}}{1 + e^{1+\sum_{j=j}^{K} \alpha_j GS_{ij}}}$	$\alpha_j \sim Normal(0, 2)$	

Table S2: Simulation parameters for gTADA. These parameters were from previous studies (De Rubeis, et al., 2014; Nguyen, et al., 2017)

	Main parameter	Other values of parameters to test gTADA power
Trio numbers	5122	1000, 2000, 10000, 25000
Case numbers	404	
Control numbers	3654	
DN mean relative risk 1 ($\bar{\gamma}_{dn1}$)	24.6	
DN mean relative risk 2 ($\bar{\gamma}_{dn2}$)	3.71	
CC mean relative risk ($\bar{\gamma}_{cc}$)	4.44	
α_0	-3.1 (~ 835 risk genes)	
ρ_{cc}	0.66	
ν_{cc}	1947	

Table S3: Type I error rates for enriched gene-set identification. The second column is obtained by setting p values < alpha thresholds and low CI > 0 while the third column is for p values < alpha thresholds. The last column is the percentage of gene sets having low CI > 0.

Alpha level	Type I error rate		
	Low CI > 0 and p value < alpha	P value < alpha	Low CI > 0 (%)
1.00E-04	5.51E-04	5.51E-04	1.6
2.00E-04	9.52E-04	9.52E-04	1.6
5.00E-04	1.70E-03	1.70E-03	1.6
1.00E-03	2.90E-03	2.90E-03	1.6
1.00E-02	1.22E-02	1.43E-02	1.6
2.00E-02	1.49E-02	2.47E-02	1.6
2.50E-02	1.53E-02	2.92E-02	1.6
3.00E-02	1.55E-02	3.43E-02	1.6
5.00E-02	1.58E-02	5.55E-02	1.6

Other tables are in SupTables (SupTable_gTADA.xlsx)

		Sup Table Name	Sheet Name
Table	S4	All gene sets used in this study	FullGeneSet
	S5	Gene-set (GS) results from gTADA	GeneSetResults
	S6	Prioritized genes for all disorders from gTADA (based on gene sets)	pGenesFromGSs
	S7	GTEx-tissue results from gTADA	GTExResults
	S8	Prioritized genes for all disorders from gTADA (based on GTEx tissues)	pGenesFromGTEx
	S9	Drug-target gene-set results from gTADA	DrugTargetResults
	S10	Drug-class gene-set results from gTADA	DrugTargetResults
	S11	Prioritized genes for all disorders from gTADA (based on drug-name gene sets	pGenesFromDrugTarget
	S12	Prioritized genes for all disorders from gTADA (based on drug-class gene sets)	pGenesFromDrugClass
	S13	Genetic parameters of EPI in de novo + case/control model.	EPI_geneticPars
	S14	Case/control ratios of EPI population samples.	EPI_CCratioUseBootstraping
	S15	GeNets enrichment results	GeNetsEnrichment

De Rubeis, S.*, et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 2014;515(7526):209-215.
Epi K. consortium and Epilepsy Phenome/Genome Project. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. *Lancet Neurol* 2017;16(2):135-143.
Hu, Y.*, et al.* Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS genetics* 2017;13(6):e1006836.
Nguyen, H.T.*, et al.* Bayesian Integrated analysis of multiple types of rare variants to infer risk genes for schizophrenia and other neurodevelopmental disorders. *bioRxiv* 2017:135293.